СИНТАКСИЧЕСКИЙ АНАЛИЗАТОР ПРЕДЛОЖЕНИЙ ТАТАРСКОГО ЯЗЫКА

Д.Ш.Сулейманов, А.Р.Гатиатуллин

Казанский государственный университет
Djavdet.Suleymanov@ksu.ru

Аннотация. В статье описывается подход к разработке татарского синтаксического анализатора, оперирующего структурами в виде деревьев зависимостей, и являющегося практической реализацией концептуально-алгоритмической прагматически-ориентированной лингвистической модели татарского языка на уровне синтаксиса.

Введение. Актуальность построения синтаксического анализатора (СА) татарской фразы определяется, прежде всего, двумя следующими аспектами. С одной стороны, СА это — часть фундаментальной научно-прикладной задачи описания концептуально-алгоритмической компьютерной модели татарского языка, с другой стороны — практический инструмент, позволяющий определять корректную расстановку пауз на границах словоформ, осуществлять корректную сегментацию текста и проверять правильность порядка следования слов в сегменте в составе таких программных систем, как синтезатор татарской речи, татарско-турецкий и татарско-русский переводчики, татарский корректор, разрабатываемых в совместной научно-исследовательской лаборатории Академии наук Татарстана и Казанского госуниверситета.

В настоящее время данная тема, имеющая богатую предысторию изучения для индоевропейских и ряда других групп языков [1-8], практически не исследована в аспекте построения компьютерной модели синтаксической структуры татарского языка.

Под открытой компьютерной синтаксической моделью языка мы понимаем концептуально-структурную модель синтаксиса языка, реализованную на компьютере и отражающую всевозможные синтаксические формулы, являющиеся корректными с точки зрения определенных норм, допустимых в естественном языке (ЕЯ).

Синтаксический анализатор можно определить как функциональную компьютерную систему, реализованную на основе

концептуально-структурной модели, и достигающей наибольшей эффективности за счет использования прагматически-ориентированных технологий разработки лингвистических моделей [9].

Синтаксическую структуру некоторой фразы на ЕЯ Ю.Д.Апресян и др. [1] представляют как дерево зависимостей, т.е. как ориентированное дерево, в узлах которого стоят определенные варианты лексико-морфологического разбора слов фразы, а дуги помечены именами синтаксических отношений. Кроме такого «классического» представления, исследователями активно используются и другие способы описания модели синтаксиса фразы, такие как, грамматики с ограничениями (constraint-based grammar) [7], модифицированные расширенные сети перехода Вудса [6], двухкомпонентная модели синтаксиса [5], грамматика непосредственных составляющих [3] и др.

Синтаксическим анализом называется построение синтаксических структур, соответствующих данной фразе, в той или иной нотации. В данной статье описывается подход к построению СА, оперирующего структурами в виде деревьев зависимостей.

На данном этапе разработка СА осуществляется при следующих ограничениях, упрощающих анализ:

- орфографическая корректность предложения и соответствие его структуры литературным нормам татарского языка;
- предложение простое;
- в предложении отсутствуют однородные члены;
- в предложении отсутствуют вводные слова;
- длина аналитических неделимых конструкций не более трех словоформ.

В качестве базового материала по синтаксису татарского языка использованы работы М.З.Закиева [10].

1. Разработка программного комплекса

Синтаксический анализатор представляет собой систему программных модулей, на вход которой поступает предложение на татарском языке, а на выходе получается результат синтаксического разбора в виде вариантов деревьев зависимости. Структурная схема синтаксического анализатора представлена на рис.1.

Рис.1. Структурная схема синтаксического анализатора

Рассмотрим каждый из составляющих СА.

1.1. Морфологический анализатор (МА) — это блок программных модулей, на вход которых поступает словоформа татарского языка, а на выходе получаются варианты разложений данной словоформы на морфемы с указанием части речи выделенной основы. По аналогии с определением, приведенным в работе [2], варианты морфологического разбора будем называть омонимами.

Пример.

Пусть на вход МА поступает слово: *балаларга 'детям / рожать'*

На выходе получается следующая таблица (Таб.1):

Таблица 1

№	Основа	Аффиксы	Часть речи
1	*бала 'ребенок'*	*-ЛАр+ГА*	сущ.
2	*балалау 'рожать'*	*-ЫргА*	глаг.

Функциональная схема данного МА и его реализация подробно описаны в работе [11].

1.2. Определитель аналитических неделимых конструкций (ОАНК) — это программа, которая на входе получает список омонимов, подаваемых как результат работы МА. На выходе ОАНК формирует список омонимов, включающий возможные аналитические неделимые конструкции.

Пример.

- Пусть на вход МА поступает исходное предложение: *Малай урманга чаклы барган. 'Мальчик шел до леса'.*

На выходе МА формирует следующую таблицу (Таб.2):

Таблица 2

№	Основа	Аффиксы	Часть речи
1	*малай 'мальчик'*		сущ.
2	*урман 'лес'*	*-ГА*	сущ.
3	*чаклы 'до'*		неиз.
4	*бару 'идти'*	*-ГАн*	глаг.

Соответственно, на выходе ОАНК, получается следующая таблица (Таб.3):

Таблица 3

№	Основа 1	Аффиксы 1	Основа 2	Аффиксы 2	Часть речи
1	*малай 'мальчик'*				сущ.
2	*урман 'лес'*	*-ГА*	*чаклы 'до'*		сущ.
3	*бару 'идти'*	*-ГАн*			глаг.

Как видно из данного примера, в данном предложении ОАНК определил одну аналитическую неделимую конструкцию (АНК): *урманга чаклы 'до леса'.*

В силу начального ограничения ОАНК можно определять аналитические конструкции длиной до трех словоформ, например такие, как:

бара торган иде — 'бывало ходил'

барган булырга тиеш — 'должен был пойти'

Определение аналитических неделимых конструкций производится путем сопоставления пар соседних омонимов. В случае успешного результата проводится дополнительное сопоставление данной пары с третьей словоформой в целях проверки — не является ли это сочетание конструкцией из трех словоформ?

В ходе сопоставления ОАНК не всегда способен однозначно определить: является ли данное сочетание аналитической неделимой конструкцией. Поэтому на выход подаются все возможные варианты.

Пример:

- На входе МА имеем: *Малай барып карады. ('Мальчик попробовал сходить / Мальчик пошел и посмотрел')*

На выходе МА получается следующий результат (Таб.4):

Таблица 4

№	Основа	Аффиксы	Часть речи
1	*малай 'мальчик'*		сущ.
2	*бару 'идти'*	*-Ып*	глаг.
3	*карау 'смотреть'*	*-ДЫ*	глаг.

На выходе ОАНК формируется следующая таблица (Таб.5):

Таблица 5

№	Основа 1	Аффиксы 1	Основа 2	Аффиксы 2	Часть речи
1	*малай*				сущ.
2	*бару*	*-Ып*	*карау*	*-ДЫ*	глаг.
2	*бару*	*-Ып*			глаг.
3	*карау*	*-ДЫ*			глаг.

Первый вариант представляет единую конструкцию:

барып карады — 'попробовал сходить'.

Второй вариант представляет конструкцию, где одна словоформа зависит от другой:

барып карады — 'пошел и посмотрел'.

ОАНК производит определение пяти типов аналитических конструкций:

1) сочетание словоформы с послелогом или послеложным словом,

2) сложный глагол в виде сочетания существительного или прилагательного и глагола,

3) сложный глагол в виде сочетания нескольких глаголов или глаголов с предикативными словами,

4) фразеологизмы,

5) имена собственные.

Для определения сочетания словоформ с послелогами или послеложными словами имеется словарь послелогов и послеложных слов с указанием контекста применения данного послелога, т.е. после какого аффикса может следовать данный послелог. Соответствующий пример приведен в таблице 6.

Таблица 6

Аффиксы	Послелоги	Примеры
-ГА	*таба*	*урманга таба 'в сторону леса'*
-ГА	*чаклы*	*урманга чаклы 'до леса'*
[]	*чаклы*	*урман чаклы 'величиной с лес'*

Особенностью первого типа сочетания являетя то, что обе словоформы фиксированные друг относительно друга. Это означает, что если в словоформе заменить последний аффикс на другой, то данная конструкция может перестать быть АНК.

Для определения конструкции второго типа, т.е. сложного глагола, образуемого путем сочетания существительного и вспомогательного глагола, используется словарь сложных глаголов.

Пример фрагмента словаря приведен в таблице 7:

Таблица 7

Основное слово	Вспомогательное слово	Перевод
вафат 'мертвый'	*булу 'стать'*	*'помереть'*
гаеп 'вина'	*итү 'делать'*	*'обвинять'*
хәбәр 'сообщение'	*итү 'делать'*	*'сообщать'*
уйга 'в мысль'	*чуму 'нырять'*	*'задуматься'*

Особенностью данного типа сочетаний является то, что первая словоформа является фиксированной, несущей лексический смысл, а вторая свободно-изменяемой, т.е. может принимать любые формы и служит для выражения грамматических признаков конструкции.

Например:

- *хәбәр итә 'сообщает',*
- *хәбәр итте 'сообщил',*

- *хәбәр итәчәк 'сообщит'.*

Для определения третьего типа сочетаний словоформ используется словарь вспомогательных глаголов с указанием правила их применения с основным глаголом (см. таблицу 8).

Таблица 8

Аффикс спереди	Вспомогательное слово	Пример
-Ырга	*тору*	*барырга тора 'хочет пойти'*
-ЫЙ	*тору*	*бара тора 'продолжает идти'*
-Ып	*тору*	*барып торды 'регулярно ходил'*

Для четвертого и пятого типов сочетания словоформ производится поиск в словаре этих типов сочетаний.

1.3. Построитель деревьев зависимостей (ПДЗ) — это совокупность модулей устанавливающих всевозможные соответствия между составляющими предложения.

Структура ПДЗ представлена на рис.2.

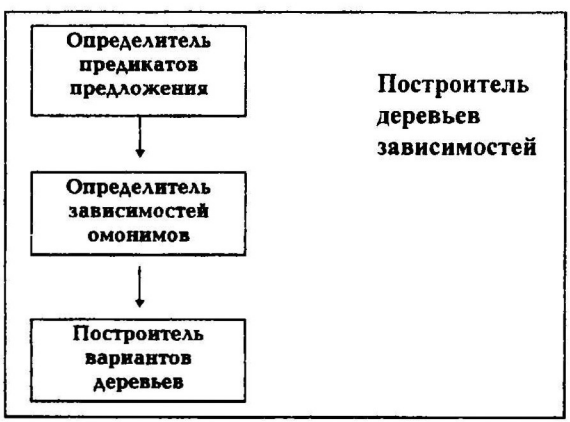

Рис.2. Структурная схема построителя деревьев зависимостей

Далее рассмотрим каждый из блоков, представленных на рис.2.

а) Определитель предикатов предложения (ОПП)

Определение предикатов основывается на условии, что в синтаксически корректном предложении предикат находится в конце предложения. Если в конце предложения расположены аналитические неделимые конструкции, то может быть найдено несколько вариантов предикатов.

Пример:

Пусть на вход МА поступает предложение: *Малай барып карады. 'Мальчик попробовал сходить'/ 'Мальчик пошел и посмотрел'*

На входе ОПП формируется следующая таблица (таб.9):

Таблица 9

№	Основа 1	Аффиксы 1	Основа 2	Аффиксы 2	Часть речи
1	*малай*				сущ.
2	*бару*	*-Ып*	*карау*	*-ДЫ*	глаг.
2	*бару*	*-Ып*			глаг.
3	*карау*	*-ДЫ*			глаг.

Следовательно, на выходе ОПП определяются два предиката:

1) *бару+Ып карау+ДЫ 'попробовал пойти',*

2) *карау+ДЫ 'посмотрел'.*

б) Определитель зависимостей омонимов

Определитель зависимостей омонимов (ОЗО) представляет собой программный модуль, формирующий для каждого омонима список ссылок на те омонимы, от которых он может зависеть.

Построение ОЗО основывается на том синтаксическом правиле, что зависимый омоним в татарском предложении располагается левее главного. Поэтому для каждого омонима указываются ссылки только на те омонимы, которые стоят в предложении правее от него. Никаких ссылок на другие омонимы предикаты не имеют.

Для определения возможных зависимостей используется словарь типов сочетаний, сформированный на основании типов сочетаний, приведенных в работах М.З.Закиева [10]. Данный словарь имеет следующую структуру (таб.10):

Таблица 10

Зависимый омоним			**Главный омоним**	
Часть речи	**Аффиксы**	**Послелог**	**Часть речи**	**Аффиксы**

Приведем пример фрагмента данного словаря (таб.11):

Таблица 11

Зависимый омоним			Главный омоним		Пример
Часть речи	Аффиксы	Послелог	Часть речи	Афф.	
Сущ.	*-ГА*		глаг.		*Укучыга әйтү 'сказать ученику'*
сущ.	*-ДАн*		глаг.		*Кешедән оялу 'стесняться человека'*
суш.	*-ДА*		глаг.		*Урамда йөрү 'ходить на улице'*
сущ.	*-ЧА*		глаг.		*Дусларча сөйләшү 'говорить по дружески'*
сущ.	*-ДАй*		глаг.		*Көзгедәй тигезләнде 'выровнилось словно зеркало'*

Процесс определения зависимостей происходит следующим образом:

- каждый омоним сопоставляется с каждым омонимом, расположенным в предложении правее него;
- для каждой сопоставляемой пары в словаре типов сочетаний ищется пара с соответствующей структурой;
- если поиск завершился успешно, то к списку ссылок левого омонима добавляется ссылка на правый омоним.

Однако проверки, основанные только на словаре, приводят к порождению большого числа ссылок, по которым строятся также и неправильные деревья зависимостей.

Так, например, следующее правило, приведенное на таблице 12, является справедливым только при непосредственном соседстве

омонимов в предложении (т.е. если это изафет). Поэтому, если отбирается данное правило, то далее производится дополнительная проверка на непосредственное соседство с целью отсечения некорректных вариантов деревьев.

Таблица 12

Часть речи	Аффиксы	Послелог	Часть речи	Аффиксы
Суш.			Сущ.	

Пример:

Пусть на вход СА подается предложение: — *Кара яңа өскәтергә түгелде. 'Чернила пролились на новую скатерть.'*

Если не производить дополнительной проверки, то ОЗО сформирует такой набор ссылок, на основании которого возможен также следующий вариант дерева зависимостей (а) (рис.3):

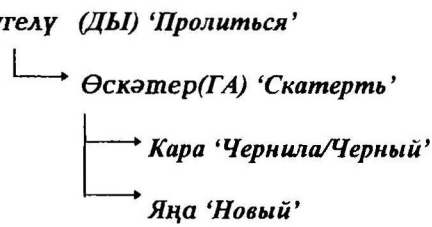

Рис.3. Дерево зависимостей (а)

Этому дереву соответствует перевод: *'Пролилось на новую черную скатерть'.*

Утверждение1. В предложении с двумя главными членами существительное без аффиксов может зависеть от существительного с одним из падежных аффиксов только в том случае, если в этом предложении имеется хотя бы еще одно существительное без падежных аффиксов.

Например, в предложении *Малай йортка керде 'Мальчик зашел в дом'* существительное *малай 'мальчик'* согласно Утверждению 1 не может зависеть от другого существительного, поскольку иначе такое предложение будет интерпретироваться как предложение без подлежащего (т.е. для данного примера некорректно). Однако, следуя

данному утверждению нельзя избежать ситуации неоднозначности, которая получается при анализе предложений следующего типа:

Карт агач йортка керде 'Старик зашел в деревянный дом'

Здесь СА выдает следующие варианты деревьев зависимостей (Рис.4 и Рис.5):

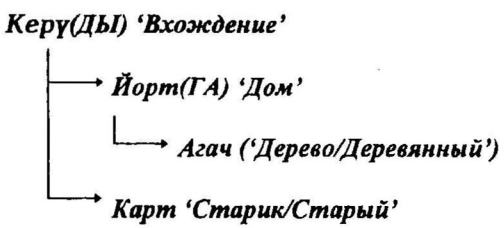

Рис.4. Дерево зависимостей 2

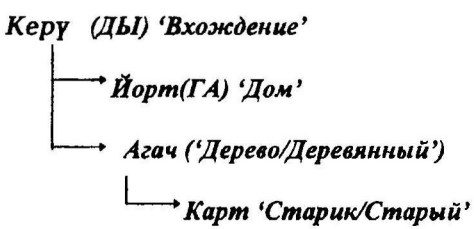

Рис.5. Дерево зависимостей 3

Дереву на рис.4 соответствует перевод: *'Старик зашел в деревянный дом.'*

Дереву зависимостей на рис.5 соответствует перевод: *'Старое дерево зашло в дом.'*

в. Построитель вариантов деревьев зависимостей (ПВДЗ) — это программный модуль, на вход которого поступает список омонимов. Для каждого из них указан набор ссылок на те омонимы, от которых он может зависеть. На выходе получаются все возможные варианты деревьев зависимостей с данными ссылками.

Алгоритм работы ПВДЗ представляет собой алгоритм перебора с возвратом, когда строятся всевозможные варианты деревьев по данным ссылкам и на выход подаются те из них, которые содержат в качестве узлов дерева омонимы всех словоформ предложения.

Рассмотрим данный алгоритм:

1) берется предикат и для него находятся все зависящие от него омонимы, которые и составляют узлы дерева первого уровня;

2) для каждого из узлов первого уровня находятся зависимые от него омонимы, которые и образуют второй уровень узлов дерева зависимостей;

3) процесс продолжается до тех пор, пока либо не закончатся все словоформы предложения, либо ни одну из оставшихся омонимов нельзя будет присоединить к построенному дереву;

4) если в полученное дерево вошли все словоформы исходного предложения, то построение считается успешным и данный вариант дерева подается на выход;

5) затем начинается формирование следующего дерева, где в одном из омонимов, имеющего несколько ссылок, используется ссылка, которая ранее не использовалась;

6) процесс продолжается до тех пор, пока не будет испробована возможность построить деревья со всеми имеющимися ссылками.

2. Программная реализация синтаксического анализатора

Синтаксический анализатор представляет собой программный модуль, включающий в себя описанные выше блоки и словари. СА может быть включен в состав других программных продуктов, для чего достаточно подключить основную процедуру программы, на вход которого подается предложение для разбора. Поэтому СА может не иметь своего собственного интерфейса. Однако для отслеживания работы программного модуля в ходе разработки и отладки создан пользовательский интерфейс с визуальными компонентами, который представлен на рис.6.

Интерфейс программы отладки представляет собой диалоговое окно с несколькими панелями, содержащими визуальные компоненты. Три средние панели отбражают результаты работы трех основных блоков синтаксического анализатора:

- данные, получаемые на выходе морфологического анализатора;
- данные, получаемые на выходе определителя аналитических неделимых конструкций;
- данные, получаемые на выходе построителя деревьев зависимостей.

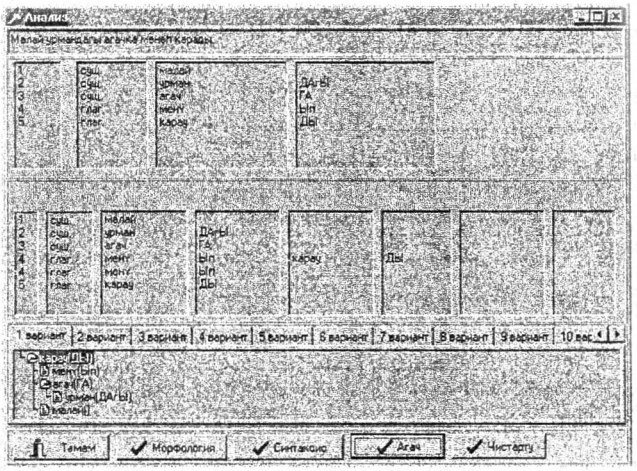

Рис.6. Интерфейс программы отладки

Рассмотрим элементы интерфейса, размещенные на диалоговом окне, показанном на рис.6.

В верхней части окна расположено поле для ввода исходного предложения (Рис.7):

Рис.7. Поле для ввода предложения для разбора

Следующая панель содержит набор списков для отображения результатов морфологического анализа введенного предложения (Рис.8):

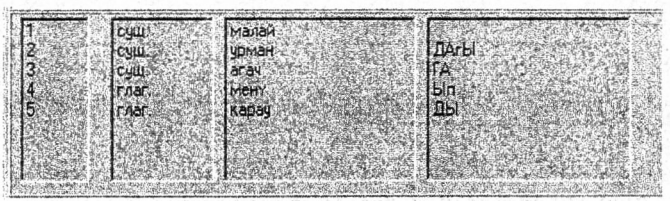

Рис.8. Окно с результатами работы МА

Как видно из рис.5, на этой панели:

- первый столбец содержит номера разбираемых словоформ,

- второй столбец содержит части речи к которым относятся, полученные омонимы,

- третий столбец содержит (омонимы) корневые морфемы — как результат работы МА,

четвертый столбец содержит наборы аффиксальных морфем морфемы — как результат работы МА.

Третья панель содержит списки, получаемые на выходе ОАНК (Рис.9):

- первый столбец содержит номера вариантов омонимов или аналитических конструкций;

- второй столбец содержит указание частей речи омонима или главного члена в АНК;

- третий, пятый и седьмой списки содержат корневые морфемы омонимов или членов АНК;

- четвертый, шестой и восьмой списки содержат последовательности аффиксов омонимов или членов АНК.

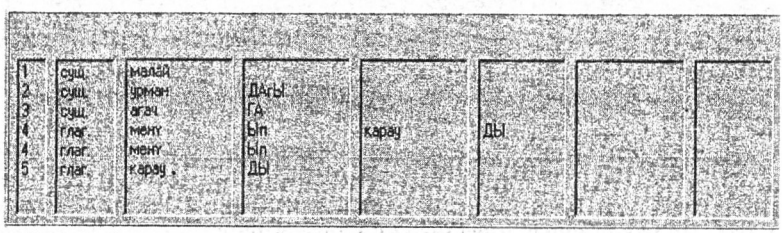

Рис.9. Окно с результатами ОАНК

Четвертая панель содержит многостраничный объект, на каждой странице которого расположены варианты деревьев зависимостей (Рис.10):

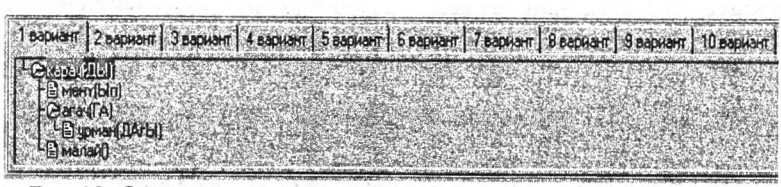

Рис.10. Окно с вариантами деревьев зависимостей

В нижней части окна расположены кнопки, активизирующие каждый из основных блоков СА (Рис.11):

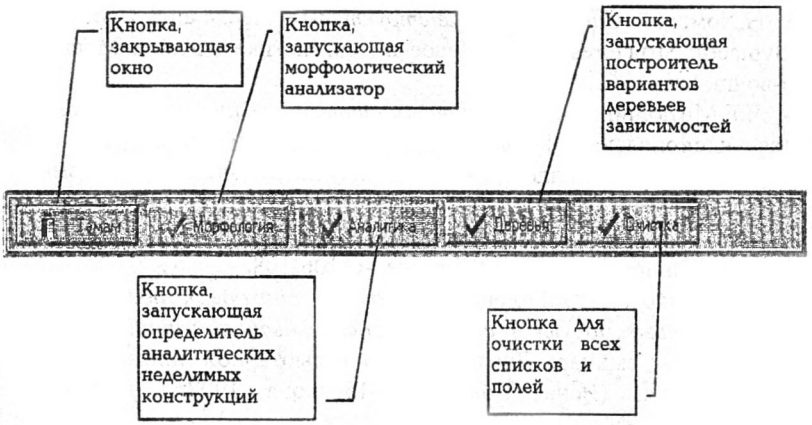

Рис.11. Кнопки для операций

Программа реализована в системе Delphi для работы в операционной системе Windows' 95 и выше.

Заключение. В данной статье описано создание модели синтаксического анализа татарского языка с использованием известного метода построения деревьев зависимостей.

Разработка синтаксического анализатора в настоящее время находится на начальном этапе и будет продолжена по пути уточнения вариантов деревьев зависимостей с минимазацией некорректных вариантов. Также по мере развития СА будут последовательно сниматься ограничения на структуру предложения, описанные во введении.

ЛИТЕРАТУРА

1. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. *Лингвистический процессор для сложных информационных систем.* - М.: Наука,1992. -256с.

2. Н.В.Григорьев. *Восходящий алгоритм построения дерева зависимостей для системы ЭТАП-3* // Труды межд. семинара Диалог'99 по компьютерной лингвистике и ее приложениям (Таруса, 1-6 июня, 1999г.). Под ред. А.С.Нариньяни. -М., 1999.

3. Н.Хомский. *Синтаксические структуры* // Пер. с англ.: *Chomsky N.* Syntactic Structures в сб. «Новое в лингвистике», вып.2. Москва: Изд-во ин.лит., 1962.

4. Л.Г.Митюшин. *О многозначных синтаксических структурах* // Пятая национальная конференция с международным участием «Искусственный интеллект-96» / Сб.научн.трудов в трех томах (Казань, 5-11 октября 1996г.). Т.1. - Казань, 1996. -С.109-114.

5. Э.В.Страхович. *Двухкомпонентная модель синтаксического анализа и машинный перевод* // Там же. -С.106-108.

6. И.А.Волкова, И.Г.Головин. *Об одном подходе к построению синтаксического модуля в системе распознавания устной речи* // Труды Межд. Семинара Диалог'97 по компьютерной лингвистике и ее приложениям (Ясная Поляна, 10-15 июня, 1997г.). Под ред. А.С.Нариньяни. -М., 1997. -С.61-62.

7. Müürisep. *Syntactic analysis of Estonian using constraint Grammar* // Труды Межд. Семинара Диалог'98 по компьютерной лингвистике и ее приложениям в двух томах. Т2. Под ред. А.С.Нариньяни. -Казань, 1998. -С.619-625.

8. Z. Güngerdü, K.Oflazer. *Parsing Turkish using the Lexical-Functional Grammar formalism.* Machine Translation, 10:293-319, 1995.

9. Сулейманов Д.Ш. *Обработка ЕЯ-текстов на основе прагматически-ориентированных лингвистических моделей* // Сб.под ред. Соловьева В.Д.: Обработка текста и когнитивные технологии. - Вып.3. Труды научного семинара "Когнитивное моделирование" (Пущино, октябрь 1998 г.). -Пущино, 1998. -С.205-212.

10.Закиев М.З. *Татарская грамматика.* Т3. Синтаксис. -Казань: Таткнигоиздат, 1992. -488 с.

11.Сулейманов Д.Ш., Гатиатуллин А.Р. *К разработке Лемматора татарских словоформ* // В сб. Трудов Международного семинара Диалог-99 по компьютерной лингвистике и ее приложениям в двух томах, г.Таруса, 31 мая -4 июня 1999, Т.2. -1999. -С.306-314.